Unspoken Words Recognition: A Review

Manish Kumar Saini, J. S. Saini, Neeraj
Electrical Engineering Department
Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana, India
Email: neeraj_bagari@yahoo.com

***Abstract* - In recent years, unspoken words recognition has received substantial attention from both the scientific research communities and the society of multimedia information access networks. Major advancements and wide range of applications in aids for the speech handicapped, speech pathology research, telecom privacy issues, cursor based text to speech, firefighters wearing pressurized suits with self contained breathing apparatus (SCBA), astronauts performing operations in pressurized gear, as a part of communication system operating in high background noise have propelled words recognition technology into the spotlight. Though early words recognition techniques used simple maximum likelihood algorithms only but the recognition process has now graduated into a science of mathematical representations and comparison processes. This survey paper provides an up-to-date review of the existing approaches and offers some insights into the study of unspoken words recognition. A number of typical techniques and EMG based approaches are discussed in this paper. Furthermore, a discussion outlining the incentives for using recognition techniques, the applications of this technology, and some of the difficulties plaguing the current systems with regard to this topic have also been provided.**

***Keywords*-Speech Pathologies, Electromyography, Hidden Markov Models, Myoelectric Signals (MES).**

I. Introduction

Unspoken words recognition relates to the task of enabling speech communication in the absence of acoustic signal. In this technique, data is acquired from elements of the human speech production process such as articulators, their neural pathways, or the brain itself. It produces a digital representation of speech which can be synthesized directly, may be interpreted as data, or fed into a communication network. Persons who have undergone a laryngectomy, or older citizens for whom speaking requires a substantial effort, would be able to mouth words rather than actually pronouncing them. For this, the unspoken words recognition has proved as an aid [2]. Alternatively, those unable to move their articulators due to paralysis could produce speech or issue commands simply by cognitively concentrating on the words to be spoken. Further, as SSI (Silent Speech Interface) is build upon the existing human speech production process, augmented with digital sensors and processing, they have the potential to be more natural sounding, spontaneous, and intuitive to use than such currently available speech pathology solutions as the electrolarynx, tracheaoesophageal speech (TOS), and cursor-based text-to-speech systems [1].

With reference to the block diagram of Fig 1, the facial EMG signal is taken by different speech articulators, which are responsible for the production of speech sounds by different data acquisition techniques as discussed in Table-I. Signal acquired is then normalized and activity detection of each signal is done [4]. From the normalized signal, features are extracted [6], [7]. After feature extraction, data analysis and feature selection is done [9], [10]. These selected features are fed into different type of classifiers [12], [14] and comparison between spoken and unspoken is done [23], [24].

The content of this article includes Introduction, generalized block diagram, comprehensive analysis of the references which provides an easy distinction between different techniques of SEMG (surface electromyography), conclusions and future perspectives.

II. Related Work

Unspoken words recognition is a medium of speech communication without using the sound when people tend to vocalize their speech sound. It is a type of electronic reading of facial muscles by the computer, identifying the phoneme and words that an individual attempts to pronounce from non-auditory sources of information about their speech movements. If the input to such a computer based system is plain text that is it does not contain additional phonems, the information system is called text-to-speech system. By creating synthetic model of human physiology, articulatory speech synthesis is accurately possible. Experimental system has evolved seven different types of technology which are used for acquiring the speech signal [2]. Substantial improvements in the word recognition have led to the development of additional sensors like throat microphones which are used as part of multimodal speech recognition. Experimental studies results in adequate growth of computing power and minimizing the electronic size which reduces the impact of noise [3]. Application field of SSI comprises assistance to a person who has undergone laryngectomy, an alternative to the electrolarynx, to oesophagal speech and tracheo-oesophagal speech. Due to its non-invasive property, SEMG provides good time resolution in clinical applications as it is well adopted in imaging and analysis [4].

EMG based technology has an interesting property that the little or acoustic energy produced during speech can also be detected. EMG activity can be detected even when the subject whispers and moves the mouth without producing the sound [7], [13]. EMG to speech approach is preferable in human to human communication particularly when there is no restriction of vocabulary and direct mapping of EMG signal is allowed to collect required speech content [15], [16]. Automatic speech recognition system is inherently ro-

DOI: 03.LSCS.2013.3.24

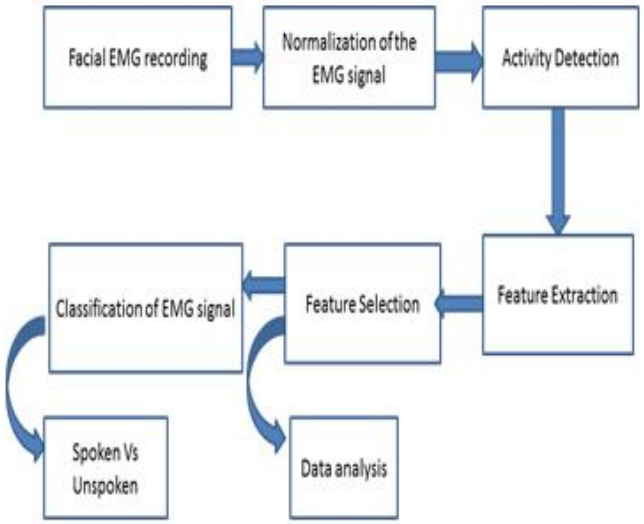

Figure 1: Generalized block diagram of unspoken word recognition

bust to high background noise, as the EMG electrodes measure the muscle activity of the skin tissue and is not based on transmitted signal in the air, so it allows confidential input in public places and provides robustness to ambient noise [17]. EMG measures the bio potential in terms of electric current that is generated by the muscle during its contraction which describes its neuro-muscular activities and these movements of the muscles are translated into speech signal [18].

In past two decades, SEMG has attracted more and more applications in rehabilitation and human computer interface [18]. The potential applications of SEMG are for facial motion disorders such as stroke patients may have swallowing disorders, for paralytic patients who suffer from illness or accident and many prosthetic device are also controlled by SEMG [20], [21]. Therefore, SEMG provides a valuable reference in clinical diagnosis and biomedical application. A special neural feature of electromyography is that, movements are not only

TABLE I. COMPREHENSIVE ANALYSIS OF REFERENCES

Sr.	Authors	Data collection specification methods and sampling frequency	Methods of feature extraction\selection	Classification technique	Efficiency/Efficacy
1	Robin Hofe et al. [5]	Magnetic sensing by employing electro-magnetic articulography Sampling Frequency: 50Hz to 500Hz	Acoustic signal is reconstructed by using Mal Frequency Cepstral Cofficient (MFCC). HMM	Viterbi alignment algorithm	An approach to mitigate the influence of additional noise can be overcome by employing a set of additional pellets.
2	J.M.Gilbert et al. [6]	Magnetic sensors are placed on the tongue and lips of subject and record the movement of the two. Sampling Frequency: 500-1000 samples/sec	Wavelet Transform(WT)	Statistical modeling technique	Very high recognition rate can be achieved with reduced sets of implants to achieve better performance and preserve the acoustic signal from distortion.
3	Bradley J.Betts et al. [7]	Facial EMG is captured by using Ag/Agcl electrodes as a sensing device Sampling frequency:10000samples/sec	Mximum likelihood algorithm, WT	Neural network classifier	Consistency of computational requirement necessitates to typically work in wearable environment. Real time performance of the system to be quantified.
4	K. Yoshida et al. [8]	For high recognition performance HMM recognition method is used to efficiently treat contextual and allophonic variation which utilizes acoustic knowledge.	Hidden markov model(HMM), WT	single Gaussian probability density function	For high recognition rate a neural network classifier can be employed and additional noise can be minimized.
5	H. Manabe et al. [9]	Three channels surface electromyography is used for acquiring the training data. Sampling Frequency: 1000samples/sec.	Multi-stream HMM	Mal frequency cepstral coefficient (MFFC)	By employing five channels EMG recognition accuracy may be improved.
6	Erik.J.Scheme et al. [10]	Five channel MES data is collected by using Ag/Agcl duotrode bipolar electrode pairs placed over the five articulatory muscles of the face. Sampling frequency:5KHz	HMM	maximum likelihood output of viterbi algorithm	Feasibility and accuracy of the real time acoustic data using its automatic segmentation & expansion of phoneme.
7	Noboru Sugie et al. [11]	Using three channels of EMG and Ag/Agcl electrodes raw EMG signals are collected. Sampling frequency: 1250samples/channel	Linear Discriminant Analysis (LDA) for discrimination of vowels		Co articulation may cause difficulty so continuous voice production should be there.
8	Ki-Seung Lee et al. [12]	A 3-channel EMG with Ag/Agcl surface electrodes are used for data acquisition Sampling frequency:100msec length of hamming window is used to extract features at each 20msec/ interval	Continuous hidden markov models, MultivariateGaussia n distribution.	Artificial Neural network	For optimal performance, the data collection must be large and recognition rate can be improved.
9	Quan Zhau et al. [14]	Five channel MES data is collected by using Ag/Agcl duotrode bipolar electrode pairs placed over the five muscles of face. Sampling frequency:10kHz	Principal component analysis, MFFC	Gaussian mixture model	Automatic myoele-ctric can be developed to make the existing system more robust.
10	S.P.Arjunan et.al [19]	SEMG is used to measure the relative activities of four facial muscles. Window size of 20 samples corresponds to 10msec	Root mean square	Error Back Propagation Neural Network	Future possibility may include speech based computer control in high background noise.

DOI: 03.LSCS.2013.3.24

voluntary but also have emotional control which leads to the hypothesis that these voluntary controls are mediated by integrated network [22].

Focusing on natural and existing EMG based speech recognition systems, the most practiced approaches are phoneme recognition, vowel recognition and complete words recognition [25], [26]. The main disadvantage of existing speech interfaces is their limited robustness in the presence of high background noise, so several electro-myographic approaches have been developed in which acoustic speech recognition is replaced by automatic speech recognition [2]. These approaches overcome the ambient noise and also produce an alternative human computer interaction for the persons having speech disability [10]. SEMG generates the relevance of multi modal interfaces which are used as a way of communication and reducing the information load in human-human and human-agent systems. In space, both input and output are limited due to severe conditions of atmosphere and acoustic signal is the most convenient way of communication [27], [28]. The purpose of this article is to present a light expression facial EMG in which electrodes work as a sensor, placed on the specific muscles needed to be examined and develop a recognition system.

III. Conclusions

Unspoken words recognition needs to be evaluated on purely "silent" databases in which sound is neither vocalized nor whispered. Specification of articulators will clearly reveal silent speech. Another improvement that could be expected in the field of words recognition is that there may be increase in training data and an adaptive learning scheme could be employed for improving the system performance. By enabling the addition of new words to the word based system and investigating the sources of noise such as electromagnetic waves and effects of gravity on speech signal, the system performance can be improved. Optimal location of electrodes should be known so there is a requirement of detailed analysis between the function of each facial muscle and the words spoken.

References

[1] S. P. Arjunan, D. K. Kumar and W. C. Yau, "Unspoken Vowel Recognition Using Facial Electromyogram", Proceedings of the IEEE EMBS Annual International Conference, New York City, USA, pp 2191-2194, 2006.

[2] B. Denby, B. Schultz and T. Honda, "Silent speech interfaces", Journal of Speech Communication, vol.no 52, pp 270–287, 2010.

[3] C.Jorgensen and D.Lee, "Sub auditory speech Recognition Based on EMG/EPG signals", In Proc., Joint Conf. on Neural Networks, vol. no 4, pp 3128–3133, 2003.

[4] T. Hueber and G. Chollet, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips", Journal of speech communication, vol.no.52, pp 288-300, 2010.

[5] R. Hofe, S. Stephen and M. Fagan, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing", Journal of Speech Communication, 2012.

[6] J. Gilbert, S. Rybchenkoa, R. Hofeb, S. Ell and R.K. Mooreb, "Isolated word recognition of silent speech using magnetic implants and sensors", Journal of medical engineering and physics, vol. 32, pp 1189-1197, 2010.

[7] B. J. Betts, K. Binsted and C. Jorgensen, "Small-vocabulary speech recognition using surface electromyography", Interacting with Computers, vol.18, pp 1242–1259, 2006.

[8] K. Yoshida, T. Watanabe and S. Koga, "Large vocabulary word recognition based on demi-syallable Hidden markov model", IEEE Transaction, C&C Information Technology Research Laboratories, pp 1-4, 1989

[9] H. Manabe, A. Hiraiwa and A. Sugimura, "Unvoiced speech recognition using EMG—mime speech recognition", Proc. Conf. Human Factors in Computing Systems, pp 794–795, 2004.

[10] E. J. Scheme, "Myoelectric signal classification for phoneme based speech recognition", IEEE Transaction on Biomedical Engineering, vol.no.54, pp 694-699, 2007

[11] N. Sugie and K. Tsunoda,"A speech prosthesis employing a speech synthesizer vowel discrimination from perioral muscle activities and vowel production", IEEE Transaction on Biomedical Engineering, vol. no.32, no.7, pp 485–90, 1985.

[12] Ki-Seung Lee,"EMG-based speech recognition using Hidden Markov Models with global control variables", IEEE Transaction on Biomedical Engineering, vol.no. 55, pp 930-940, 2008.

[13] Wai Chee Yau, S.P Arjunan, D. K. Kumar, "Classification of voiceless speech using facial muscle activity and vision based Techniques", School of Electrical and Computer Engineering, vol no.4, pp 1-4, 2001.

[14] Q. Zhou and N. Jiang,"Improved phoneme-based myoelectric speech recognition", IEEE Transaction on Biomedical Engineering, vol.no.56, , pp 2016-2023 , 2009.

[15] S. P. Arjunan, D. Kumar and K. Wheeler, "Spectral properties of surface EMG and muscle conduction velocity: A study based on SEMG Model", Proceedings of School of Electrical and Computer Engineering, pp 23-26, 1998.

[16]M. Janke, M. Wand and K. Nakamura,"Further investigation on EMG-to-speech conversion**",** IEEE Cognitive Systems Lab, ICASSP, pp 365-368, 2012.

[17] S. C. Jou and T. Schultz, "Automatic speech recognition based on electromyographic biosignals", Cognitive Systems Laboratory Karlsruhe University, Karlsruhe, Germany ICASSP, 2012.

[18] Md. R. Ahsan, "EMG signal classification for human computer interaction: A Review", European Journal of Scientific Research ISSN 1450-216, vol. no.3, pp 480-501, 2009.

[19] S. P Arjunan, D. K. Kumar and W.C. Yau, "Vowel recognition of english and german language using facial Movement for speech control based HCI", Conferences in Re-search and Practice in Information Technology (CRPIT), Vol.56, 2006

[20] A.J.Fridlund and J.T.Cacioppo, "Guidelines for human electromyographic research", Journal of Biological Psychology, vol. 23, pp 567-589,1986.

[21] Cheng-Ning Huang, "The review of applications and measurements in facial electromyography", Journal of Medical and Biological Engineering, vol.no.25, pp 15-20, 2004.

[22] G. Lapatki, D. F. Stegeman, and I. E. Jonas, "A surface EMG electrode for the simultaneous observation of multiple facial muscles", Journal of Neuroscience, vol. 123, pp 117-128, 2003.

[23] K. Englehan, B. Hudgins and M. Stevenson, "Classification of the myolectric signal using time-frequency based represenlations", Medical Engineering and Physics, vol.no.21,

DOI: 03.LSCS.2013.3.24

pp 431–438, 1999.

[24]G. Gibert and M. Pruzinec, "Enhancement of human computer interaction with facial electromyographic sensors", OZCHI Proceedings ISBN, pp 23-27, 2009

[25] Kumar, S., Kumar, D.K., Alemu, M, Burry, M., "EMG Based Voice Recognition", Proceedings of Intelligent Sensors, Sensor Networks and Information Processing Conference, pp 593–597, 2004

[26] E. Lopez, M.Javier and Javier Minguez, "Syllable-Based Speech Recognition Using EMG", 32nd Annual International Conference of the IEEE, PP 4699-4702, 2010.

[27]C. DeLuca, "Physiology and mathematics of myoelectric signals", IEEE Transaction on Biomedical Enggineering, vol.no.26, pp 313–325, 1979.

[28] K. Binsted and C. Jorgensen, "Sub-auditory speech recognition", ICS Department, University of Hawaii, 2001.

DOI: 03.LSCS.2013.3.24